Draft for Comments

Chapter: Context and Content

Text Book on

Fundamentals of Multimedia Computing

Authors:
Gerald Friedland
and
Ramesh Jain

Draft for Comments

Chapter: Context and Content

Introduction

Human perceptual system has been a topic for active exploration by philosophers for long time. Almost two century ago [Ber] George Berkeley asked: "**If a tree falls in a forest and no one is around to hear it, does it make a sound?**" Sound is defined as the sensation excited in the ear when the air or other medium is set in motion. Thus, if there is no 'receiving ear' than there is no sound. In other words, perception is not only data – it is a close interaction between the data, transmission medium, and the interpreter. This is shown in figure 1.

Multimedia communication and computing is fundamentally related to the perception problem. In any perception problem, there are three components that must be considered.

- The data acquired for an environment
- The medium used to transmit physical attribute to the perceiver
- The perceiver

Characteristics of each of these must be considered in designing and developing a multimedia system. It has been very well realized, and rigorously articulated and represented, that we understand the world based on the sensory data that that we receive using our sensors and the knowledge about the world that we have accumulated since our birth [Popper, Neisser]. Both the data and the knowledge are integral component of the understanding

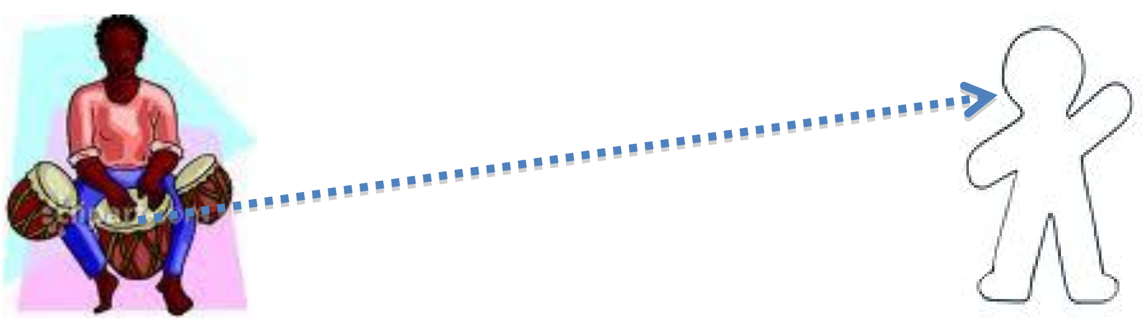


Figure 1: Perception requires a source, medium and a perceiver.

Let us revisit multimedia computing from fundamentals. Where does the multimedia data come from? Why do we even need multimedia? What is the multimedia content problem?

Multimedia data, such as visual (photos and video), aural, and other sensory data are captured for an event that unfolds over time. Each medium represents a particular physical spatio-temporal attribute of the event. An event represents changing

relationships among objects and these are captured by different media. The data captured by any kind of sensor really represents these spatio temporal physical attributes of the environment. Objects are part of the environment and their physical attributes are also captured by the media.

Each sensor only captures one type of physical attribute from its ‘perspective’ afforded from its physical location, including its orientation. Multiple sensors could be combined to create the composite data representing a synchronized signal obtained from these sensors. Thus, one uses appropriate number and types of sensors to capture all attributes of the event that may be of interest in a particular application. Multimedia is the right approach to capture event information and experiences because each medium captures only one physical attribute and taken as a whole, the multimedia stream is capable of combining the correlated and complimentary information from individual streams to provide more holistic information and experience than possible using any one medium. None of the individual medium, including the most powerful human senses (the vision), can capture holistic experience in most applications. Humans have five senses and combine them to experience events in real world.

Equally important is the fact that each sensor captures data about the environment from its position and perspective. If its position or perspective is changed, then the data and experience also change. For interpretation of the data, one must know the position and perspective. Moreover, many sensors, like cameras, have several other parameters (focal length, aperture diameter, flash, etc.) that determine the capture of the data and hence they are very important in understanding and analyzing the experience represented by the data.

Most important component in multimedia computing systems is the user. Each user is unique and while interacting with a system, the context may be different. Interpretation of the data is not only user dependent but also dependent on the context of the user. If you give the same photo to different people and ask them to assign tags to represent the photo, there may be as many different tags as the number of people assigning tags. Moreover, many studies have demonstrated that if you give the same photo to the same person at two different times in different contexts, then the tags assigned are different. The concept of Rorschach tests is based on the theory that an interpretation of data is as much, or more, dependent on the person than the data.

Connecting Data and Users

Multimedia computing addresses a problem that many other fields like computer vision, databases, and information retrieval face: connecting data and users. As shown in figure 2 below, data exists in many forms, ranging from bits to alphanumeric documents to photos and video. On the other hand users of the data in a modern computing environment may come from many different education backgrounds, of different culture, and of different socio-economic status. The challenge is how to connect a user with a data source so the user can use the data he needs to solve his application. A key point to

remember is that a user is never interested in what and where the data is; she is only interested in solving the problem at hand.

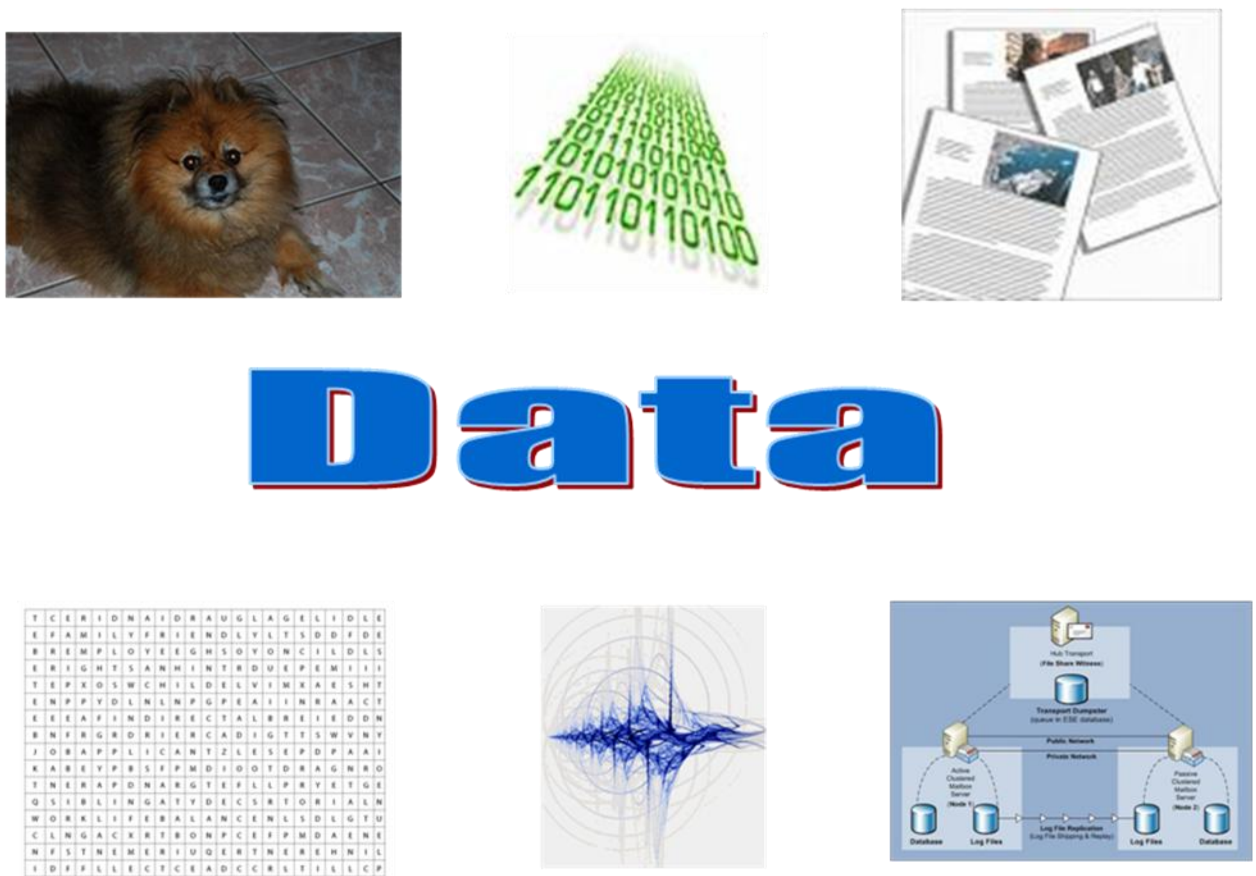


Figure 2: A user normally combines information from multiple sources to form a holistic concept about an object or event.

The major hurdle in connecting users to the data is often referred to as the semantic gap, as stated in [11]:

“We opine that most of the disappointments with early retrieval systems come from the lack of recognizing the existence of the semantic gap and its consequences for system set-up. *The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a*

given situation. A linguistic description is almost always contextual, whereas an image may live by itself."

To understand semantic gap, let's consider figure 3. This figure shows that the data operations in a computer start at the bit level and can be structured to represent various data concepts such as documents, images, and videos. A user on the other hand always starts thinking in terms of objects and events and builds other concepts based on the basic notion of objects and events. The transformation of data level concepts such as images to user level concepts such as objects and events is the challenge that must be solved by content analysis and used in content organization and retrieval.

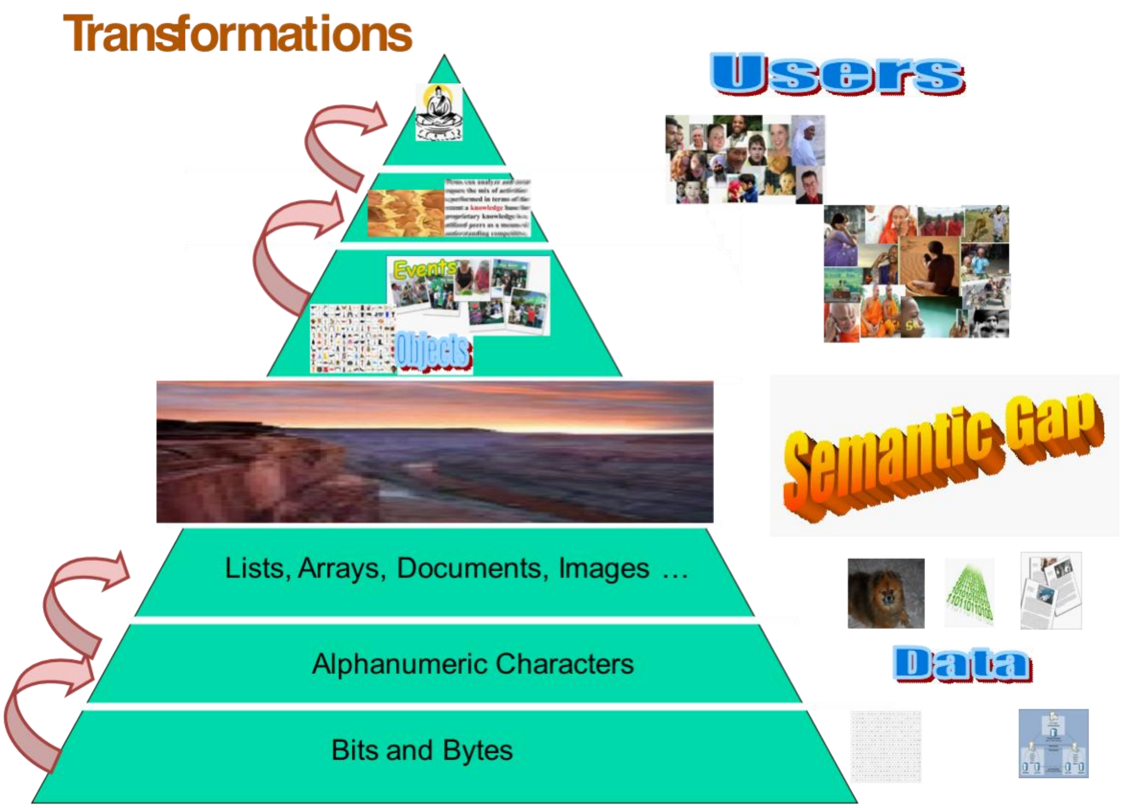


Figure 3: Semantic gap is created because computer operates using bits and structuring them into bytes and other structures. Humans think in terms of events and objects and build their concepts in terms of those. How do we go from bits to objects and events is the gap that must be bridged to allow smooth interactions between humans and computers.

Content and Context

Content usually refers to data values in a multimedia stream, such as images or audio. It is commonly understood that we refer to substantive information perceived by a user in the data that is represented by a particular file. Content analysis and content-based retrieval represents majority of research done in multimedia and related fields like

computer vision. Thus a photo may contain a person standing near a car next to a house. The challenge faced by content analysis is the famous problem of pattern recognition. In all sensory data, the problem is to segment the data into meaningful parts and to use known models of objects of interest to label all segments of an image. This is where one runs into a tricky situation: We need segments to recognize objects but we also need objects to segment the data. It is possible to formulate this problem such that one can use models of potential objects to segment and then see how best the segments fit object recognition. One may potentially use an optimization framework to accomplish this. The problem gets complicated and almost intractable because in some cases, most notably in images, a higher dimensional space is mapped into 2-dimensional space resulting in loss of information making the problem impossible to solve unless some strong assumptions are made. This is the reason behind the sensory-gap [11] defined as: *The sensory gap is the gap between the object in the world and the information in a (computational) description derived from a recording of that scene.*

In many sensors the signals from multiple objects get intermingles and combined making it almost impossible to solve the problem. The only way to simplify the problem appears to be to use other information and reduce the number of potential objects that could be in the data and and then check which objects are most likely to result in the signal.

Philosphers and scientists have been trying to solve the mystery of human perception for several centuries and are still far from any seemingly right solution. Closer to multimedia, people have been working on image recognition and speech recognition (notice only speech recognition, not audio recognition) and are still far from being close to solving these problems even with the powerful computing infra-structure that we have. The successful solutions usually are for limited domains meaning the number of objects is limited in those applications, making the problem more tractable.

Let us look at a related concept: context. Context is defined in standard dictionaries and reference sources as:

- The set of circumstances or facts that surround a particular event, situation, etc.
- The interrelated conditions in which something exists or occurs: environment or Setting.
- Determinant of meaning.

In technical areas, context started receiving attention in the last decade and has been receiving increasing attention. The precise definition of context is very difficult because content and context are closely related, as one could easily see from the dictionary meaning of context. The role of context in understanding our environment has been emphasized by several researchers in psychology of perception, neurophysiology, and cognitive sciences. Clearly, a perceiver is at least as important as the data in perception. The perceiver uses extensive context in understanding signals and recognizing objects and events in it.

When using multimedia, each media data stream may act as context for other. Consider a video. It has an image sequences and an audio stream. As is easily obvious, without

audio we can not fully understand video and without visual signal we can not fully experience video. In multimedia, each media stream has incomplete information about the environment captured and must utilize complementary information in some other media stream to recreate the environment to understand and enjoy it.

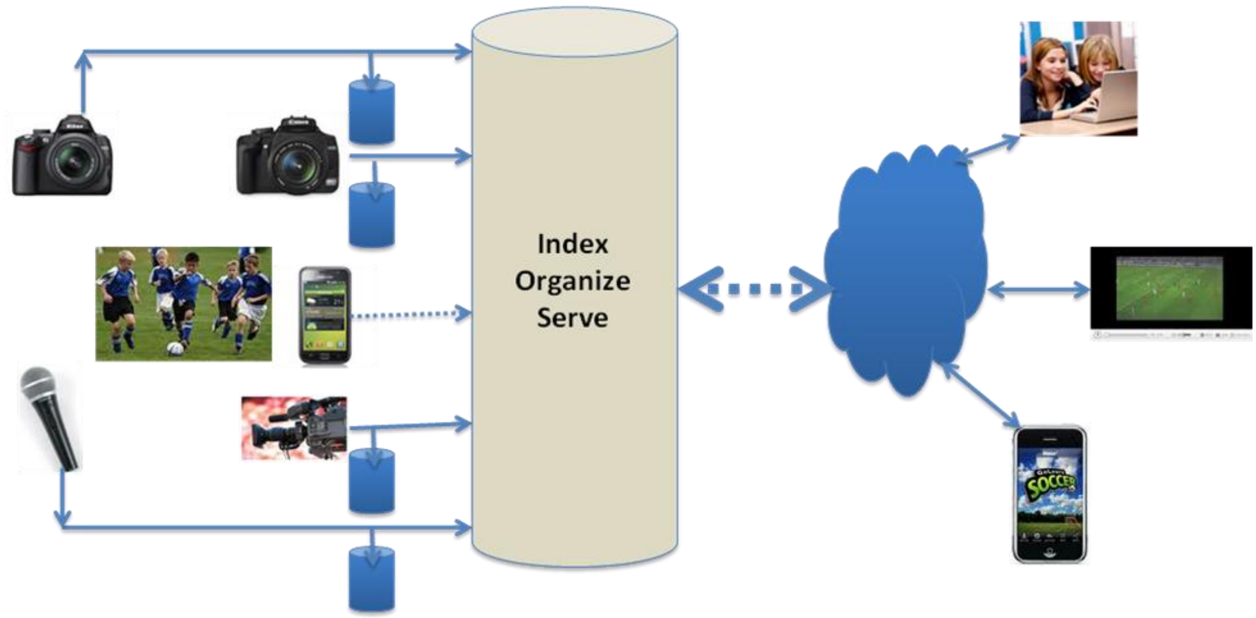


Figure 4: An eco-system showing creation to consumption of multimedia data. Capture depends on the location, environmental conditions and device parameters. All data is locally stored and collectively organized and indexed for serving consumers on different device.

Types of Context in Multimedia Systems

The context or knowledge that could be used in analyzing data may be considered in following different classes:

1. **Context in Content:** Relationship among different objects and even in their subparts in real world can be converted to relationships among data items and can be applied in analysis of data.
2. **Device Parameters:** Environmental parameters of the digital devices at the time of data collection play important role in the analysis of data.
3. **Data Acquisition Context:** Knowledge about the person collecting data, location, and environmental conditions at the time of data acquisition (sun angle, cloudy, rainy, night, indoor, etc.) affects the content and knowledge of these could be used in data analysis.
4. **Perceiver:** Cognitive scientists know the importance of the perceiver. Rorschach tests are a clear demonstration of the knowledge and personality of the perceiver in interpretation of visual data.

Interpretation Context: Real world situation in which the data is interpreted results in focus on different aspects of data. A botanist looks at the garden with different goal and interprets it differently than a person interested in enjoying the beauty of flowers.

Consider a simple case to understand how context can significantly simplify analysis: A photo is taken and needs to be interpreted. If one knows when the photo was taken and at that time what was the lighting level in the scene, one could use appropriate parameters for segmentation and interpretation of images. Moreover, if the photo was taken in Iowa, one should not expect beaches or mountains.

In the following, we discuss the role of each of these types of context. Explicit consideration of the above classes of contexts helps analysis and management of multimedia data. In the following we show some examples of these contexts.

Context in Content

In image understanding, one starts with the intensity values at every pixel and computes edges, regions, and other features. These features are then used to find meaningful objects and events in the image. The role of context in image understanding, commonly called computer vision, was emphasized even in very early days by systems developed for the block worlds and those for indoor scenes like MYSYS. Waltz [4] developed a constraint propogation framework to label junctions and understand configuration of blocks. This work was one of the earliest work in computer vision and the one that also resulted in starting to explore role and propagation of constraint.

A very clear role of context was demonstrated in 1976 in MYSYS system [8], where it is stated: "In scene analysis, it is frequently impossible to interpret parts of an image taken out of context. Different objects may have similar appearances, while objects belonging to the same functional class can have strikingly different appearances (e.g., chairs). Ambiguous local interpretations must be ruled out by using contextual constraints to achieve a meaningful, globally consistent interpretation of the whole scene."

Further they said:
"Scene interpretation is an attempt to explain observed sensory data in terms of prior knowledge about the depicted domain. The explanation can entail many types and levels of knowledge, some of which may be probabilistic or inconsistent. It must also allow for the likelihood that the data is noisy. For these reasons, scene interpretation is not a purely deductive problem with a unique correct solution; it is a problem that requires a search for the best or optimum explanation."

Early research in scene understanding resulted in realization that problems in scene understanding are posed by providing (1) a set of possible assignments for each region, with associated a priori likelihoods for each assignment and (2) a set of constraints that are derived for the types of scenes to be analyzed and that determine the a posteriori likelihood of regions.

The above approach, resulted in popular framework commonly called relaxation labeling, has been formalized and used in many computer vision systems. The basic idea in this approach is to utilize knowledge of local relationships among objects that may appear in a scene and use these local relationships to propagate local interpretations repeatedly to refine overall interpretations

over the image. Thus one may use simple facts like 'a computer monitor is on a desk', 'floor is at the bottom in an image', and 'desk is on the floor' in an office scene. A set of such constraints among all objects can then be used iteratively in the interpretation process. The process starts with recognition of all possible regions and assigning them all plausible levels. The relaxation process then iteratively eliminates all implausible levels. When this process terminates, each region is assigned best possible interpretation based on the constraints, or the knowledge, available.

Relaxation processes have been defined at pixel levels. Also the relaxation processes have been developed considering a set of labels such that each iteration eliminates certain labels. Another approach uses probabilistic assignment of labels by updating the probability of labels in each iteration.

Verification Vision

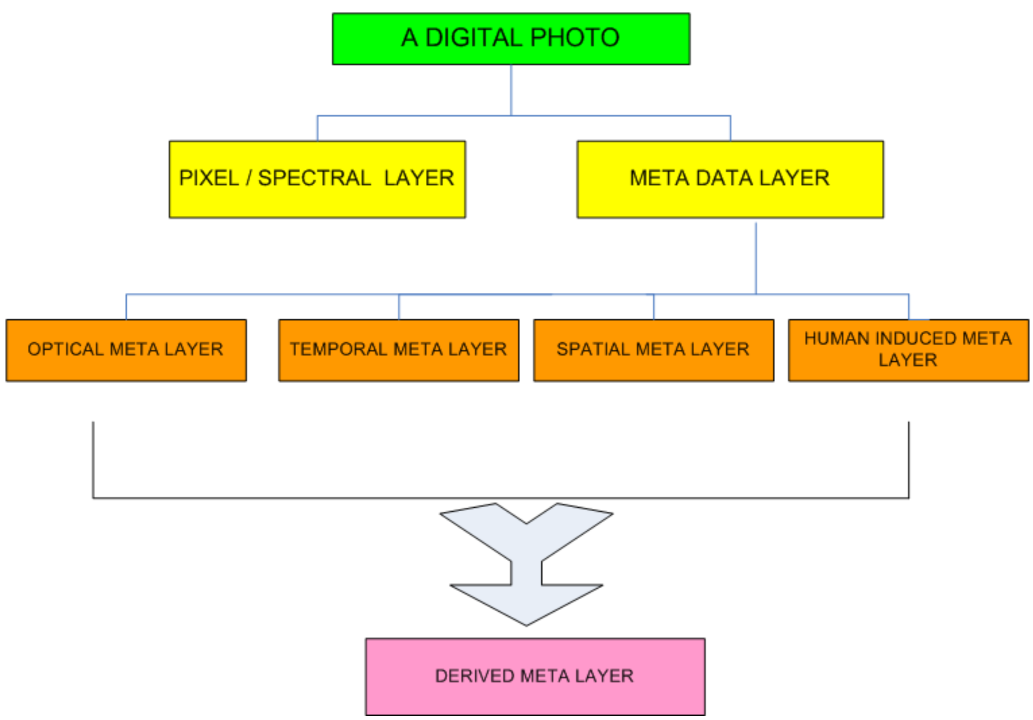
Another commonly used approach for recognizing objects using context is called verification vision. Suppose that we are given an image of an object and we need to find how many times and where this object appears in an image. Such a problem is essentially a verification, rather than an object recognition, problem. There are many approaches for verification. One of the most popular one is called template matching. A verification algorithm can be used to exhaustively verify the presence of each model from a large model-base, but such an exhaustive approach will not be a very effective method. A verification approach is desirable if one, or at most a few, objects are possible candidates. If one is given some contextual information then that could be used to constrain number of potential objects in an image. Thus, if one knows that an image was taken inside the kitchen of a western home, then one can list number of objects that may need to be verified; clearly an elephant or a tall banyan tree will not be in that list.

Context Only Image Search: Commercial Systems

The most commonly used example of use of context in search are the commercial application of image search from any major search engine, Google, Bing, or Yahoo. Suppose that you search for images with keyword Obama, rose, Tendulkar, or cars. Look at the results of the image search and you will be surprised. Most of the results are correct. Surprisingly, as is well known, most of these results are obtained without even opening the image file – that is without even looking at the content. These search systems only use the context provided to them from sources such as the name of the file containing picture, surrounding text on the page where the picture file appears, and the topic of the page. These search engines perform much better than any content based retrieval system that we have seen, including the ones that one of the authors was involved in developing.

Device Parameters

For considering a common example of device parameters, lets consider a device that most of us use regularly: a digital camera. Unlike their predecessors, modern digital cameras are more 'event capture' devices rather than photographic devices. Modern digital cameras have multiple sensors, including GPS to determine the location of the photograph, present on it that can capture much more information about the photo shooting event. The digital photograph is no longer just a collection of pixels; it has a host of sensory information stored as metadata. This information can be summarized if we model the digital photograph as a multilayered structure as shown in Figure 5. The layers are: i. Pixel/ Spectral layer and ii. Meta Layer. The pixel layer stores information recorded by the CCD (as pixel values). The content of the image is present in this layer. The Meta Layer stores the contextual information about a photo shoot. The Meta Layer can further be divided into the following sublayers: a. *Optical Meta Layer*. It contains the metadata related to the optics of the camera; e.g., the focal length,aperture, exposure time etc. These metadata store important cues about the context in which the image was shot (like the lighting condition, depth of field and distance of subjects in the image). b. *Temporal Meta Layer*. It contains the time stamp of the instant in which the photo was shot. The time stamp of a single image in a standalone environment is not informative enough.



Figurex 5: Different types of information stored with an image taken by a digital camera.

But in a collection of images (e.g., photo albums) the time difference can shed valuable light on the content of the images. c. *Spatial Meta Layer*. It contains the spatial coordinates of the places where pictures were shot. These coordinates are generated by

the GPS systems attached to the camera. In case of camera phones, cell-tower ids help generate this data. d. *Human Induced Meta Layer*. This layer contains the tags/ comments/ ratings posted by people. Community tagging (in online photo albums) or voice tagging in mobile phones help generate data for this layer. e. *DerivedMeta Layer*. This metadata is inferred from other information by various statistical modeling approaches. The taxonomy defined here helps us to explore the information sources present in a digital camera image. Presently, the spectral, optical and temporal layers are present in almost all digital photographs, while the spatial, human induced and derived meta layers might not be present.

The Exchangeable Image File (EXIF) Standard specifies the camera parameters recorded during a photo shoot. The actual parameters recorded depend on the particular camera manufacturer. But there are certain fundamental parameters which are recorded by all popular camera models. These are: Exposure Time, Focal Length, F-number, Flash,Metering Mode and ISO. Subject distance is an important parameter which can be used to make important inferences about the image content. However most camera models do not store this parameter.

EXIF data is attached to all digital pictures and contains very valuable information about camera parameters used in taking photos. These parameters affect the part of the scene imaged and the intensity values of pixels, while others give very valuable contextual information about the data acquisition context. For a set of 2000 images, clusters of the photos based on the Optical Metadata associated with them provide valuable insights in the role of these parameters. The following figures show the probability distribution of human induced tags in different Exif based clusters.

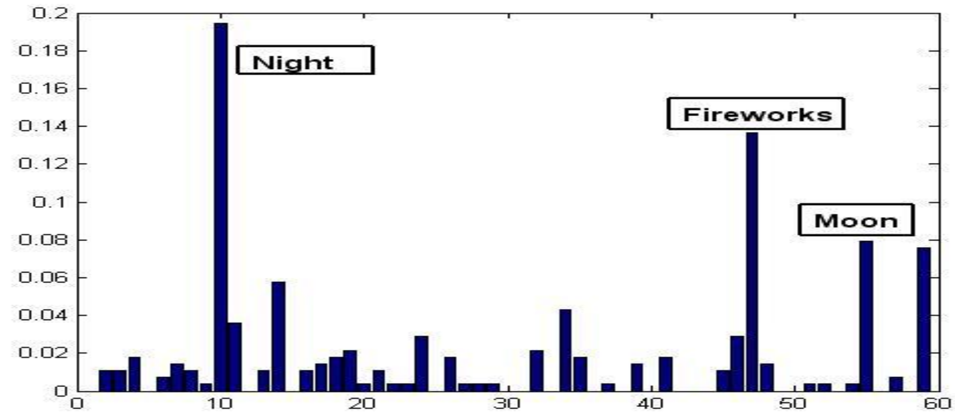


Figure 6: a cluster with No Flash and Large Exposure time

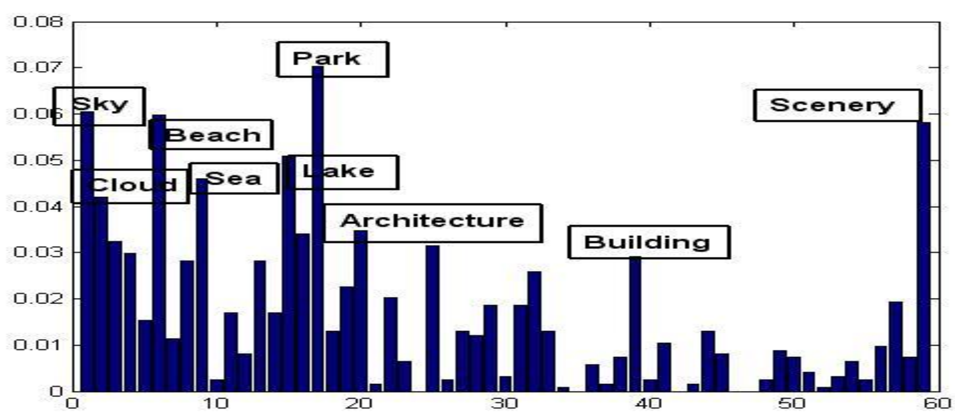


Figure 7: A cluster with NO Flash, Low Exposure and Low Focal Length (large Field of View)

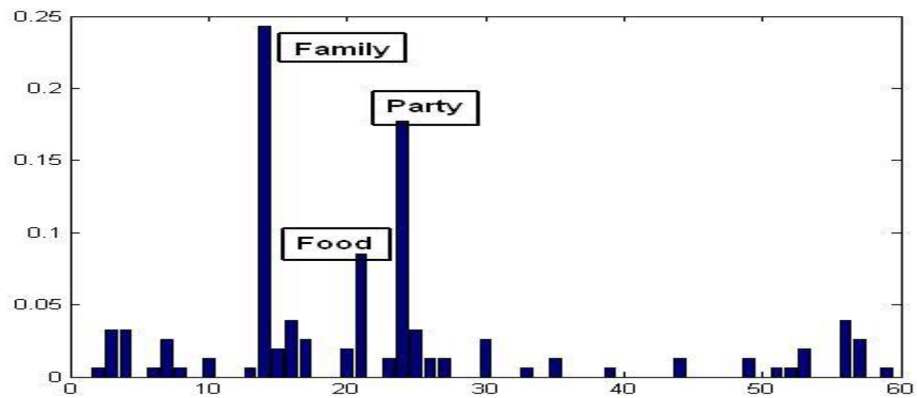


Figure 8: a cluster with Flash, Larger Exposure (mostly indoor concepts)

Sometimes the context provided by device parameters like EXIF are more meaningful than those provided by the analysis of content or intensity values. Consider the photo shown in Figure 9.

If we try to assign tags based on pixel features only, we may obtain:
Scenery, City Streets, Illuminations, Wildlife.

However, tags based on Exif features give:
Indoor, Party, Portrait, Group Photo Indoors.

Why is there a discrepancy between tags predicted by the content and the context channel? This is a Photo of a Photo. The image originally appeared in a magazine and Figure 9 is a photo of it using a standard digital camera. Since the content and context channels capture two entirely different semantics about an image, the tags are so different.

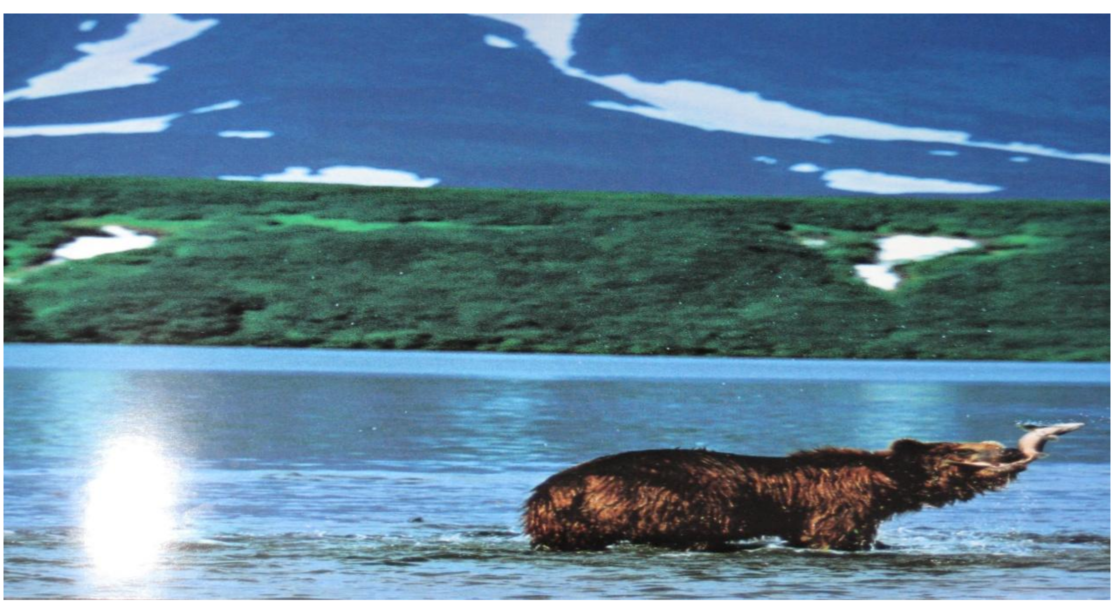


Figure 9: What do you see in this photo. Exif parameters correctly analyze that this is a photo of a photo – this is not a photo taken in the real world.

Figure 10: Where is this picture taken. Photo interpretation will mislead that this is in China, but context correctly tells that this is taken in Orlando, Florida. GPS parameters have no difficulty in putting the photo at a right place.

Perceiver

A perceiver brings general knowledge of the environment that she lives to bear on the interpretation of the data. Since we all grew up in the physical world that we inhabit, we make several contextual assumptions about the world. Our brain automatically uses these assumptions in interpretation of data. The best example of this is in visual illusions. In visual illusions, we see something that is different from objective reality. Even after measurements and knowing that what we see is not right, our cognitive system does not accept the reality. This clearly demonstrates the role of assumptions that we all use in interpretation of data. Sometimes, the interpretation of data tells a lot more about the interpreter than about the data.

Rorschach tests

Many psychological tests and even many photos that commonly appear in psychology literature show a picture to people and them to specify what they see in the picture. In these experiments, the goal is to know about the personal characteristics of the perceiver. These tests demonstrate that the interpretation of data depends on the perceiver [16]. Common phrases used by people such as 'Do you see what I see in this picture?' are clear indication of the well recognized role of the perceiver.

Interpretation Context

Domain knowledge

Look at Figure X. Try to guess what this picture is. Now suppose you are tols that this is an Atomic Force Microsope image – now you will think about either cellular images or atomic level images of materials. In fact what you are seeing is a recent breakthrough in developing chemical sensors that could be developed cost-effectively [20]. Most possibly, you and I can not understand what this is because we do not know how to interpret these images because we don't have the context and associated knowledge. But this was an important image in the announcement of this breakthrough. What this shows is that without domain knowledge, it is almost impossible for people to analyze and understand content.

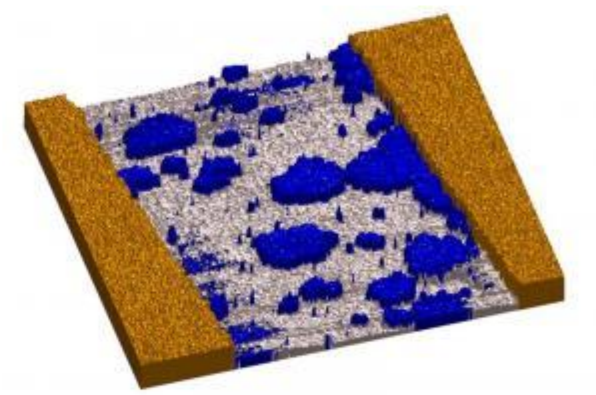


Figure 11: What do we see in this picture. It purely depends on the domain knowledge.

Context in Photo Management in SmartPhones

Next generation digital photos will be captured by smart phones. Sets. As shown in Figure 12, smartphones are privy to a lot of information sources which will help augment photos with lot of contextual knowledge. Let's consider some scenarios where context can play an useful role for organizing photos on smartphones.

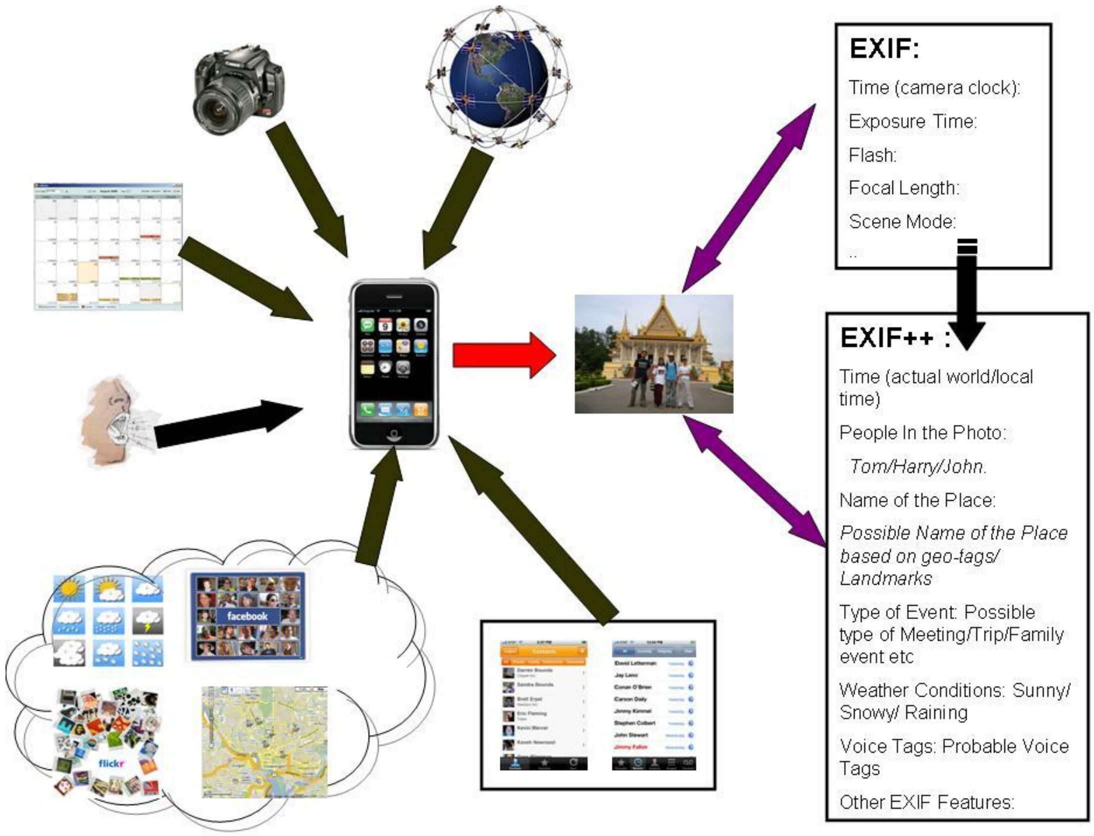


Figure 12: Showing how each photo by modern cameras will be augmented with very rich metadata – that we call EXIF++.

a. Identifying People: Most consumer cameras can detect frontal faces while shooting photos. However face recognition/assigning name tags to them is an open problem. In case of smart phones this problem can be reduced to a much easier face identification in a small set. Our personal calendars provide us with the name of people we are meeting at any particular point of time. If a portrait/group photo was shot at this time, we just need to compare with the faces appearing with the faces of people we were supposed to meet. The latter can come from any social network or photo sharing site. If we do not good match, the system can go through the list of recent callers /callee and try matching with their faces. If not we can go through our contact list to find a good matching face. This makes the problem much easier to solve.

b. Identifying objects: Object identification is another problem. Usually people shoot landmarks or special objects of interest using their smart phones. The EXIF information can tell us if the object of interest is indoors or outdoors. We can also estimate a possible size of the object based on the focal length , field of view and subject distance. Geo location will help us narrow down to a small set of important objects (e.g., landmarks or flowers or food) which are shot in that area. Comparison to this much refined set is likely to generate much better results for object identification.

c. Event name tagging based on public /private Calendars: People shoot a lot of photos in their life events, e.g., parties, trips, meeting et al. It is very relevant and useful to tag photos based on the events. It may be very difficult to automatically tag a photo with an event name (John's birthday) or even a generic class name (indoor party) based on pixels and Exif alone. However, personal calendars can help on such cases to properly tag with event names. Further, people often participate in public events like concerts, baseball games, parades etc. There are abundant sources of event repositories on the web. Based on a users location information and the events taking place in the vicinity, it may be possible to predict the proper event name (e.g., Celtics vs Lakers Game) with good accuracy.

Further Reading

In human perception, Irwin Rock [5] and Gregory[6] have strongly championed the role of knowledge in many different forms in visual perception. They believed that context plays at least as important, in most cases significantly more, role as content. Human sensory processing uses context extensively. The role of context in understanding our environment has been emphasized by several researchers in psychology of perception, neurophysiology, and cognitive sciences [4,5,6,7,8]. A review of context as used in different computing systems and environments is provided in [3, 15, 16]. All digital cameras follow a meta-data standard called EXIF to capture the context around photos taken by the camera. A detailed analysis of these parameters is provided in [19].

Many philosophers and cognitive scientists, including one of the most noted in the 20th century Karl Popper [20] and Ulric Neisser [21] have created models of all human actions that include context and prior knowledge about an application as an integral component of understanding data. Media processing techniques so far, however, have focused on content assuming that interpretation can be done based only on the data values. In semantic gap, an important point is the 'the linguistic description is always contextual'.

Context in general, and GPS in particular, has serious privacy implications. This is discussed in [18].

References

1. Sinha, P., Jain, R.: Classification and annotation of digital photos using optical context data. In: Proceedings of the 2008 international conference on Content-based image and video retrieval, ACM (2008) 309–318
2. Sinha, P., Jain, R.: Semantics In Digital Photos: A Contenxtual Analysis. In: Proceedings of the 2008 IEEE International Conference on Semantic Computing, IEEE Computer Society (2008) 58–65
3. Ye Sun Joung, Magda El Zarki, Ramesh Jain, "A User Model for Personalization Services", 3rd INTERNATIONAL WORKSHOP ON CONTEXT MODELING

AND MANAGEMENT FOR SMART ENVIRONMENTS CMMSE, November 1-4, 2009.

4. Patrick Winston: The Psychology of Computer Vision (McGraw-Hill Computer Science Series), 1975.
5. Irwin Rock: The Logic of Perception (MIT Press) 1985.
6. Richard Gregory: The Intelligent Eye (1970). London: Weidenfeld and Nicolson.
7. Ramachandran, V.S. and Anstis, S.M.: The Perception of Apparent Motion, Scientific American, June 1986.
8. H. G. Barrow and Tenenbaum, J. M.: MSYS: A System for Reasoning About Scenes, Technical Report 121, AI Center SRI, April 1976.
9. Galleguillos, C and Belongie, S.: Context Based Object Categorization: A Critical Survey.Bo Gong and Ramesh Jain, “Segmenting Photo Streams in Events Based on Optical Metadata”, Proc. IEEE Conf. on Semantic Computing, Irvine, CA, Sept. 17-19, 2007.
10. Santini, S. and Jain, R. “Beyond Query by Example; IEEE Second Workshop on Multimedia Signal Processing, Redondo Beach, CA , pp. 3 – 8, USA 7-9 Dec 1998.
11. Arnold Smeulders , et. al.,”Content-Based Image Retrieval at the End of the Early Years” IEEE Transactions on Pattern Analysis and Machine Intelligence December 2000.
12. Exner, John E. *The Rorschach: A Comprehensive System*. Vol 1: Basic Foundations. New York: John Wiley & Sons. ISBN 0-471-55902-4, 1995.
13. Freuder, E.C. Recognition of Real Objects, Vision Flash, No. 33, MIT, 1972.
14. George Berkeley, *A Treatise Concerning the Principles of Human Knowledge*, 1734. section 45
15. G. D. Abowd, A. K. Dey, P. J. Brown, N. Davies, M. Smith, and P. Steggles. Towards a better understanding of context and context-awareness. In Proc. 1st Int. Symp. on Handheld and Ubiquitous Computing (HUC’99), pages 304–307. Springer-Verlag, 1999.
16. Cristiana Bolchini, Carlo A. Curino, Elisa Quintarelli, Fabio A. Schreiber, Letizia Tanca, “A Data-oriented Survey of Context Models*”,* SIGMOD Record, December 2007 (Vol. 36, No. 4)
17. J Kittler, J Illingworth, “Relaxation labelling algorithms--a review”, Image and Vision Computing, 1985
18. G. Friedland and R. Sommer, "Cybercasing the Joint: On the Privacy Implications of Geo-Tagging", in Fifth USENIX Workshop on Hot Topics in Security (HotSec '10) at the 19th USENIX Security Symposium, Washington, D.C.
19. EXIF: Exchangeable image file format for digital cameras: Exif version 2.2. Technical report, Japan Electronics and Information Technology Industries Association, 2002.
20. Karl Popper, *Objective Knowledge: An Evolutionary Approach*, 1972, Rev. ed., 1979.
21. Ulric Neisser, *Cognition and Reality*: Principles and Implications of Cognitive Psychology Neisser, U. (1976) San Francisco: W. H. Freeman

Exercises

Why is it that at many places where they show video, such as in airplanes, they charge for renting headphones? You can see visual component of the video free but must pay for the audio component? Does this mean that audio has all information and video has none?

What is cocktail party effect? How is it that in such a noisy environment, people can carry conversations?

We need more exercises, some of them should be more clear-cut.